CHALMERS
UNIVERSITY OF TECHNOLOGY

Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy

AI & SOCIETY (2020) 35:917–926
https://doi.org/10.1007/s00146-020-00960-w

OPEN FORUM

Artificial intelligence, transparency, and public decision-making

Why explanations are key when trying to produce perceived legitimacy

Karl de Fine Licht[1] · **Jenny de Fine Licht**[2]

Received: 11 December 2019 / Accepted: 28 February 2020 / Published online: 19 March 2020

Abstract
The increasing use of Artificial Intelligence (AI) for making decisions in public affairs has sparked a lively debate on the benefits and potential harms of self-learning technologies, ranging from the hopes of fully informed and objectively taken decisions to fear for the destruction of mankind. To prevent the negative outcomes and to achieve accountable systems, many have argued that we need to open up the "black box" of AI decision-making and make it more transparent. Whereas this debate has primarily focused on how transparency can secure high-quality, fair, and reliable decisions, far less attention has been devoted to the role of transparency when it comes to how the general public come to *perceive* AI decision-making as legitimate and worthy of acceptance. Since relying on coercion is not only normatively problematic but also costly and highly inefficient, perceived legitimacy is fundamental to the democratic system. This paper discusses how transparency in and about AI decision-making can affect the *public's perception* of the legitimacy of decisions and decision-makers and produce a framework for analyzing these questions. We argue that a limited form of transparency that focuses on providing justifications for decisions has the potential to provide sufficient ground for *perceived legitimacy* without producing the harms full transparency would bring.

Keywords Artificial intelligence · Transparency · Public decision-making · Perceived legitimacy · Explainability · Framework

1 Introduction

Artificial intelligence (AI) is becoming more prevalent in every aspect of our lives. In particular, the increasing use of AI technologies and assistants for decision-making in public affairs—in taking policy decisions or authoritative decisions regarding the rights or burdens of individual citizens—has sparked a lively debate on the benefits and potential harms of self-learning technologies. This debate ranges from hopes of fully informed and objectively made decisions to fears for the destruction of mankind (e.g., Pasquale 2015; O'Neil 2016; Bostrom 2017).[1] To prevent negative outcomes and create accountable systems that individuals can trust, many have argued that we need to open up the "black box" of AI decision-making and make it more transparent (e.g., O'Neil 2016; Wachter et al. 2017; Floridi et al. 2018). This "opening up" will make it easier for us to understand (interpret) the functioning of the AI as well as possible to receive explanations for individual decisions (e.g., Zarsky 2016; Lepri et al. 2017; Zerilli et al. 2018; Binns 2018; De Laat 2018).[2]

✉ Karl de Fine Licht
Karl.definelicht@chalmers.se

Jenny de Fine Licht
jenny.definelicht@spa.gu.se

[1] Management and Economics, Chalmers Tekniska högskola AB Technology, 412 96 Göteborg, Sweden

[2] School of Public Administration, University of Gothenburg, PO Box 712, 405 30 Göteborg, Sweden

[1] The European Commission states that due to the increasing complexity of its decisions and the speed at which decisions need to be delivered, the extensive use of AI assistants is of the utmost importance: https://ec.europa.eu/futurium/en/blog/how-far-can-public-service-tasks-be-delegated-ai.

[2] This can also be seen in the first regulation on AI in Singapore. Here, transparency is a key term used to increase the public's confidence in decision-making where transparency is a core feature. https://www.mas.gov.sg/~/media/resource/news_room/press_releases/2018/Annex%20A%20Summary%20of%20the%20FEAT%20Principles.pdf. The same can be said for AI auditing agencies, such as O'Neil Risk Consulting & Algorithmic Auditing: https://www.oneilrisk.com/. There are also numerous initiatives from the industry with regard to transparency, for example: https://www.microsoft.com/

However, though a lot of interesting work has been done in the area of transparency, far less attention has been devoted to the role of transparency in terms of how those who are ultimately affected (i.e., the general public) *come to perceive* AI decision-making as being legitimate and worthy of acceptance. Researchers have noted the importance of public acceptance with regard to AI implementation (e.g., Zerilli et al. 2018; Binns et al. 2018) and there are several frameworks that can be used to make AI systems less biased, more fair, etc., (e.g., Binns 2018; Boscoe 2019; Zednik 2019) which might lead to an increase in perceived legitimacy. These frameworks etc., however, do not explicitly engage with the theories and empirical findings from the social sciences regarding how individuals' legitimacy perceptions are affected by different elements, such as the transparency of the process, decisions, or reasons behind said decisions. As well as this, no general framework exists to analyze the perceived legitimacy of AI in the broader context of the socio-technological system to elucidate the issues of today and tomorrow regarding transparency and perceived legitimacy.

This paper discusses how transparency in and with regard to AI decision-making can affect public perceptions of the legitimacy of AI decisions and decision-makers. The paper also provides a framework for transparency and perceived legitimacy in AI decision-making in the socio-technical system. This discussion is informed by relevant literature from the social sciences that is combined in a novel way to further the exploration of how perceived legitimacy can be produced in general and in the context of AI in particular. Based on our reading of the literature, we argue that a limited form of transparency that focuses on *providing justifications* for decisions has the potential to provide *sufficient grounds* for perceived legitimacy in AI decision-making. The notion that we should opt for transparency with regard to justification is not new (Binns 2018). Instead, what is new is the argument that perceived legitimacy is produced by presenting the justifications and contextualization of the mode of transparency in question, where actions and interactions are carried out by and between the decision-makers and the general public.[3]

The structure of the paper is as follows. We first explain what is meant by transparency in AI decision-making—with "AI," we refer to machine learning and deep learning algorithms as well as predictive analytics.[4] Thereafter, we discuss why transparency in AI decision-making can harm legitimacy in the eyes of the public. We then explore the potential of a justifications-focused approach to transparency and follow this with a conclusion and examination of what needs to be done in terms of future research.

2 Full transparency in AI and public decision-making

Ensuring transparency in public affairs has been widely promoted, both by policy-makers and social scientists, as a method of increasing trust and perceived legitimacy among the public (see Hood 2006). This has led to a wide range of transparency innovations, from making records publicly available on the internet to broadcasting plenary meetings. Thus, it is not strange to assume, as those involved in the debate about AI in public decision-making have done, that rendering AI decision-making processes more transparent will increase the public's trust in these processes and the decisions they lead up to.

According to common definitions, an organization or state of affairs has become transparent (or more transparent) when an actor (A) has made its workings and/or performances available (or more available) (B) to another actor (C). This can be done through various means (M). This definition is compatible with renowned definitions on transparency in the social sciences (e.g., Hood 2006; Grimmelikhuijsen 2012 for comparison) as well as those in the fields of AI and transparency (e.g., Turilli and Floridi 2009; Floridi et al. 2018). When a government (A) can make its source code available (B) to the public (C) so that they can see that nothing untoward is occurring in relation to their use of an algorithm to better predict the risk of recidivism in parole hearings, then the government has become more transparent in its processes.

Our definition of transparency arguably allows for a wide range of combinations of A, B, C, and M. Inspired by Mansbridge (2009), we argue that in relation to transparency in public decision-making, a distinction can be made between transparency that (1) informs C (e.g., the public) about final *decisions or policies*; (2) about the *process* resulting in the decisions (transparency in process); and (3) about the *reasons* on which the decision is based (transparency in rational). These forms of transparency should be understood as degrees rather than separate elements, as it is difficult to provide the reasons for a decision without making explicit what the decision is, in the same way that

Footnote 2 (continued)

en-us/research/group/fate/; https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf.

[3] Although our main focus is the use of AI technology in public policymaking, authoritative public decision-making in relation to individuals as citizens, and the intermingling of these actors (i.e., the socio-technical system), the general argument should also apply to cases of AI use in more private operations.

[4] For a thorough discussion of the definitions of AI, see Russell and Norvig 2016.

is difficult to present the process leading up to a decision without making the reasons on which it is based explicit. Thus, in most cases, the transparency of a process should be considered more transparent than the transparency of the reason, which, in turn, is more transparent than the transparency of the decision.

Intuitively, however, not all forms of transparency lead to greater perceived legitimacy. Take the classic comic segment from the show *Little Britain*, where a claimant is waiting for a decision from an official, and after the official has entered all the necessary information into her computer, she waits a moment, only to tell the claimant, "Computer says no."[5] The reason why this is so comical is mostly due to its absurdity, as it radically clashes with our expectations regarding the type of answers we should receive from officials. We expect to be treated in a way in which we can rationally accept an adverse decision, and for this we need to know the bottom-line reasoning behind the decision. In other words, we expect some insight into the decision or a certain level of transparency. However, if the official at the computer screen, instead of merely saying "Computer says no" (i.e., making only the *decision* transparent), turned her screen to show the claimant a widely inscrutable algorithm, such as a decision forest, and claimed that she has now shown the claimant the whole process, the level of absurdity would only be accentuated. Hence, in this context, this form of transparency may be a non-starter when it comes to perceived legitimacy. This example shows that with AI decision-making and its perceived legitimacy, an important question to ask is not whether we should have transparency, but rather *which kind* of transparency should be applied.

In spite of the fact that not all forms of transparency may have positive effects on perceived legitimacy, some are in favor of full transparency (i.e., both transparency in rationale and transparency in process) (e.g., Hosanagar and Jair 2018; New and Castro 2018). Assuming that the discussion regarding the implementation of AI follows the logic of public debate on transparency in general, these voices are likely to grow stronger as AI techniques develop and become more widely implemented in society. With regard to the *Little Britain* case presented above, the proponents of full transparency could say that even though the claimant does not fully understand the algorithm on the screen, it should still be made available to them, because being respectful in this way (i.e., by hiding nothing) fosters perceived legitimacy.

We believe that *if* perceived legitimacy is the goal, we should opt for transparency in rationale and not transparency in process. By transparency in rationale, we refer to the public receiving information for the justification or explanation of a decision as well as details on who can be held accountable for said decision. Thus, our meaning of transparency in rationale is similar to that of Floridi et al. (2018: 699f) and their use of "explicability," which implies that the public receives an explanation or justification for the decision made, a description of the process leading up to it, and an account of who is responsible for it. However, if explicability means that we actually make the decision-making processes fully transparent, then we do not believe it suitable in relation to the production of perceived legitimacy; but, if decision-makers should provide an explanation in the form of a narrative where it is explained how the decision has been made, then this might be applicable for perceived legitimacy.

Of course, the notion that AI assistants should be able to provide justifications or explanations for their decisions is not novel. In fact, it fits nicely with the core components of the rapidly evolving research field of explainable AI (XAI) (e.g., Gunning 2017, 2019; Thelisson et al. 2017).[6] In particular, Binns et al. (2018) have examined how different kinds of explanations affect the fairness judgments of the general public.[7] Likewise, corporations such as Google and Microsoft, as well as the Defense Advanced Research Projects Agency, are currently working toward XAI development.[8] Our discussion expands this line of reasoning by providing a more developed theoretical foundation for why explanations are critical and worthy of further exploration.

To appreciate what *full* transparency in AI and public decision-making would amount to, we propose dividing the entire decision-making process into three phases: Phase 1 is the goal-setting phase (goal-setting), Phase 2 is the coding phase (coding), and Phase 3 is the implementation phase

[5] The clip can be viewed at https://www.youtube.com/watch?v=_lu1xyYx3Eo.

[6] By XAI, we refer to the focus on what is sometimes called "subject-centric" ("explainability"), not "object-centric" ("interpretability") (e.g., Došilović et al. 2018) regarding what the public should do when they are facing a decision. AI should be able to provide reasons for why we should do something or why it did something, not an explanation of how the decision came about, assuming that this is not part of the narrative explanation. The possible explanations are varied, where natural language explanations (McAuley and Leskovec 2013), such as counterfactual explanations (Wachter et al. 2017), visualizations of learned representations, or explanations by example (Caruana et al. 1999), are just some of the most common examples. These justifications or explanations could be given beforehand (ex-ante) or afterward (ex-post), and they can be generic or specific (e.g., Wachter et al. 2017; Miller 2019).

[7] They find that the perceptions did not vary between the explanations (Binns et al. 2018). This is not relevant for our case, as we are interested in the difference between justifications and outcomes, not between explanations (or justifications).

[8] In relation to the current literature on explainable AI, however, we view the act of justifying a decision *as a form of transparency*. In addition, we argue that there may be a reasonable level of public transparency in relation to the design of AI and the decisions made by AI. See https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

(implementation) (see, e.g., de Laat: 529–533 for a comparison of AI in the market sphere and Boscoe 2019 for a similarly structured process in public decision-making). During Phase 1 (the goal-setting phase), decision-makers decide on the goals of the AI, how they should be weighed against each other when in conflict, and the features and data available to draw inferences from. For example, if you want your AI to choose which buildings you should give priority to when initiating a large renovation project, you may want it to have a feature that knows when each building was last renovated and which renovation process would be the most cost-effective to begin with. Of course, these features might pull in different directions, meaning that you will have to give the values different weights to guide the AI on what to do when said features are in conflict. These decisions are often highly political, as they require decision-makers to explicitly distinguish between advantages and disadvantages at a high level of precision.

In Phase 2 (the coding phase), the AI is developed and worked on to ensure it meets the necessary standards. This is often a point of introduction for problems related to bugs and biases, and is well described in the literature (e.g., Sweeney 2013; Datta et al. 2015; O'Neil 2016; Boscoe 2019). In this phase, it is discussed what the accuracy rates are, what they should be, how these and other performance metrics differ, how they should be allowed to differ across different subpopulations (when deciding about groups or individuals), what data to use when training the algorithm, and how to clean it. With public decision-making, the main challenge is ensuring that the AIs are *good enough* when it comes to these issues. Of course, it may be difficult or even impossible for decision-makers to know for sure whether the AIs they have authorized are up to standard without relying on programmers. This is not a problem restricted to AI decision-making, since decision-makers rely on expert opinions in virtually all policy areas. However, the problem may be accentuated in AI decision-making, as few political representatives are trained in code reading or programming. Furthermore, it may be difficult to establish goals and guarantee that their respective importance is sufficiently precise for programmers' needs.

In Phase 3 (the implementation phase), the AI is applied in the public decision-making processes, and the results produced by the AI are used in actual decision-making. This can be done by having the AI make the decision by itself or, more plausibly, by having an individual formally make the decision based on the results or recommendations of the AI. Naturally, this phase is often the one to which researchers refer when discussing AI and transparency. This phase will also, of course, feed back into Phases 1 and 2. For example, when AI assistants have been implemented in real-world settings, they are sometimes found to be discriminatory in an unintentional way and hence in need of modifications. Similarly, if left unsupervised, they might develop "bad habits" that alter the intentions of the decision-makers. Furthermore, ideological shifts among the decision-makers might require changes in goals and prioritizations, and to further complicate things, these changes among the decision-makers might be sparked by their deeper understanding of what the realization of the goals of the AI assistant would imply. Thus, there is a constant intermingling between Phases 1–3, with all phases deeply connected to each other.[9]

To make the *process* of Phase 1 (goal-setting) and 2 (coding) fully transparent to the public (C), the decision-makers (A) need to make the deliberations about goals and prioritizations as well as the deliberations of the programmers (B) available to the public, along with the training data, the testing data etc. In Phase 3, when the AI is implemented in the decision-making process, the source code and records regarding how the AI is used in the decision-making process need to be made publicly available. Ensuring transparency of the *reasons* that decisions are based on means that decision-makers should provide justifications for their decisions. This can be done in each phase (see Fig. 1). Transparency regarding the reasons will presumably contain an attempt to justify the overall functionality of the AI (e.g., its goals, their weights, its methods in different situations) (e.g., Boscoe 2019).

In the remainder of the paper, we will argue that decision-makers should in general opt for the more limited form of transparency (i.e., transparency regarding the reasons that the decision is based on, rather than transparency about the decision-making process).

3 Transparency: why more is not necessarily better

For AI in public decision-making, voices have been raised in favor of full transparency. This would mean making all three phases in the decision-making process described in Sect. 2 fully transparent. However, even though there are those in favor of full transparency, the debate regarding transparency in AI and public decision-making is currently slanted toward partial transparency (i.e., transparency in rationale), which is seen as a key element of legitimacy and perceived legitimacy. Even though we tend to agree with this, there are persuasive arguments in favor of full transparency that

[9] For an excellent and more in-depth discussion of how to make these three phases more transparent in what we mostly interpret as transparency in rationale, see Boscoe (2019). She uses what she calls "six checkpoints," where different actors at different stages should, for example, say why they do what they do so that the public can access the justification for choosing certain data, cleaning it in a certain way, etc.

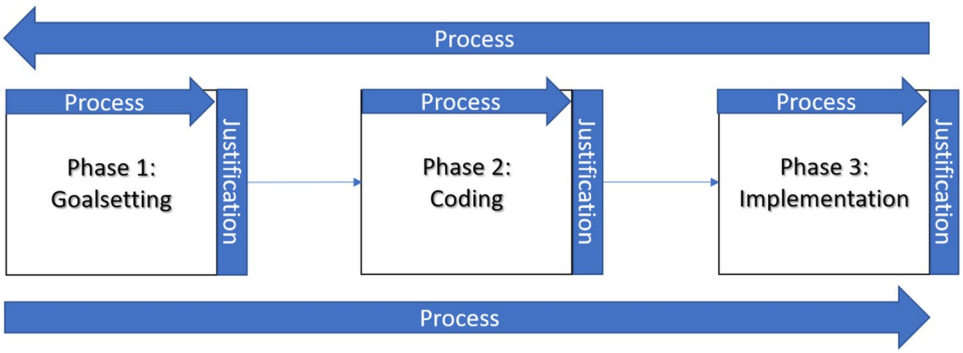

Fig. 1 The three phases of AI decision-making

have not been fully marshaled in the debate thus far. Consequently, in this section, we will present arguments in favor of why, for example, releasing the source code could lead to higher perceived legitimacy, even though the general public will not understand it. We will then evidence why we believe these arguments fail.

The first argument for rendering the decision-making process in Phases 1–3 fully transparent is that transparency makes decision-makers aware of the public eye, thereby making them aware of their responsibility to work toward the public good rather than in their own self-interests (e.g., Elster 1998; Chambers 2004, 2005; Naurin 2007).[10] This is true for both the decision-makers when establishing the goals as well as the programmers "writing the code" (i.e. choosing classifiers, test- and training data, etc.). Making them aware of the public eye could be through monitoring their meetings in real time, or through revealing the source code.

Intuitively, transparency in a process can lead to these positive effects. However, we believe that this is not the case in relation to AI in public decision-making. Beginning with Phase 3, assuming that the public needs a chance to actually understand the code, the code would likely have to be much simpler than it is in today's systems. This would arguably make it easier to game the system, and it would provide the industry with fewer incentives to innovate (e.g., Zarsky 2016; Lepri et al. 2017; De Laat 2018; Carabantes 2019).[11] Thus, in this sense, though, demands on transparency would probably produce a lower degree of quality in the decisions and thereby less perceived legitimacy.

Similarly, persuasive arguments exist against the view that full transparency would yield better decisions in Phases 1 and 2. One reason for this is that transparency about a process imposes incentives on the decision-makers to *look credible* in the eyes of the observers during their deliberations, but this is not necessarily positive in relation to the prospect of reaching a good decision. For example, Bok (1989, p. 175) argues that all organizations need

> ⋯some shelter in order to be able to arrive at choices and to carry them out. The processes of reasoning, planning, accommodation, and choice are hampered if fully exposed from the outset, no matter how great the corresponding dangers of secrecy. A tentative process of learning, of assimilating information, of considering alternatives and weighing consequences, is required in order to arrive at a coherent position.

In front of an audience, it may be difficult to change an opinion (even with a good argument), present dissenting opinions, or ask necessary but seemingly "stupid" questions (e.g., Elster 1998; Meade and Stasavage 2006; Mansbridge 2009). Asking "stupid" questions is especially important in relation to AI assistants, since these matters are so complex that there will often be many seemingly "stupid" questions to ask. Decision-makers and professionals might also become more interested in avoiding blame than in finding optimal solutions (Hood 2007), meaning that the open discussion may become shallow (Chambers 2004, 2005) and fail to produce the best possible outcomes. This suggests that making Phases 1 and 2 available is not necessarily beneficial.

Furthermore, since AI assistants work quickly, thereby producing potentially positive and negative outcomes in a short timeframe, quick and efficient decision-making may be vital when implementing AI assistants (Phases 1 and 2) and adjusting them (Phase 3). Thus, the "capacity to act" (Warren and Mansbridge 2013), or the ability to actually come to a decision and implement it, is *even more important* in AI decision-making than in regular policymaking. When the process is made transparent, however, decision-makers may be more concerned with signaling loyalty to their constituents or adhering to the special interests they represent than with finding a solution to an emerging problem. This means that important decisions may be delayed or based on

[10] It has also been argued that we might receive decisions of lower quality when deliberations in this phase are made public (De Laat 2018). We will argue in favor of this notion using a different route.

[11] For opposing views on these matters, see Carlini and Wagner (2017) and https://staff.fnwi.uva.nl/m.welling/wp-content/uploads/Model-versus-Data-AI.pdf.

an inaccurate account of the information at hand, thereby resulting in lower perceived legitimacy.

The second argument in favor of full transparency is that it should increase public *understanding* of decisions and decision-making processes, thus making the public more confident about decision-makers. There are several problems with this argument, but the most pertinent is what is sometimes called the information overload (e.g., Eppler and Mengis 2004) or the transparency paradox (Richards and King 2013). This holds that if we received all of the information from Phases 1 and 2, we would have a lot of information to assimilate, and this problem becomes even worse in Phase 3. Finding the relevant information would be as difficult as finding the proverbial needle in a haystack—something that decision-makers and the public will understand and adapt to. Hence, we will not witness the positive outcomes of making the information available. Consequently, there are strong arguments in favor of not being fully transparent but instead giving the public information about what they want and need to know to ensure that they are adequately informed about decisions.

The third argument in favor of full transparency is that it increases perceived legitimacy because it induces a *feeling of control* among the public. The idea of a close relationship between transparency and accountability underlies many discussions of transparency (e.g., Gutmann and Thompson 1996; Hood 2010; Kosack and Fung 2014). In principal–agent terminology (Holmström 1979; Fearon 1998; Manin et al. 1999), transparency allows the principal (i.e., the public) to overcome the information asymmetry regarding the agent's (i.e., the representative's) workings, thereby leading to renewed instructions about what to do or even removal from office. The reduced uncertainty is likely to render the principal more confident in delegating powers to the agent (Ferejohn 1999) and ensuring that it adheres to its decisions (i.e., regards it as legitimate).

There are reasons to believe, however, that this is not necessarily true. Transparency as a promoter of accountability can contribute to the myth of hidden politics (e.g., Fenster 2006), where the public does not believe that it actually has access to the true decision-making process. Thus, when code and data are made transparent, the public might think this release is hiding the "real code" in actual use or that there is some sort of "back door" in the code to ensure that it works in another way from what is expected. Similarly, the source code and data that the AI has been trained and tested on are often too complex for most individuals to fully understand, meaning that full transparency might actually make the public less prone to believing that they have more control over the AI than they did before they had access to the data.[12] As such, due to the staggering amount of information available, it may be difficult to understand who to hold accountable for what.

The final argument in favor of full transparency holds that transparency generates positive results regarding perceived legitimacy, as the public will perceive the decision-making processes *to be fair*, and this view will also affect their evaluations of the decisions and decision-makers (e.g., Thibaut and Walker 1975; Napier and Tyler 2008; Tyler 2010). As Gutmann and Thompson (1996, p. 95) note, according to almost every normative perspective, transparency is superior to non-transparency, and today, transparency has become a buzzword in governance. Hence, transparency in itself might have an independent positive effect on perceived legitimacy, irrespective of the content of the process that is being made available to the public. Thus, even if an individual cannot understand the functioning of a classifier (e.g., a deep neural network) or how the input generates the output, they could still appreciate the fact that the government has made the process transparent.

However, even though it is intuitive that transparent institutions are preferred over non-transparent ones and that they yield higher perceived legitimacy, recent empirical research (e.g., Grimmelikhuijsen 2012; de Fine Licht 2014) has shown that it is far from evident that increased transparency generates trust or acceptance of public policies. In some cases, the effect can even be *negative*. The problem is that full transparency reveals the actual reality of decision-making and that real-world decision-making rarely or never lives up to the democratic or professional ideal (e.g., Tsoukas 1997). In other words, the public may become disappointed when they realize that decision-making processes are, more than occasionally, characterized by a process of "muddling through" (Lindblom 1959) rather than a rational process of identifying the problems, collecting the relevant information, and carefully weighing all the alternatives. On the one hand, the recipe for such disappointments is certainly to improve the processes. On the other hand, a real-world political and private decision-making process that the public is completely satisfied with is unlikely due to the inherent conflict and inefficiencies in political affairs. In the context of Phase 3, the plausibly negative effects of making the code and data transparent can be applied here as well.

To conclude, there is much evidence to highlight that full transparency in decision-making processes does not necessarily make decision-making about or by AI legitimate in the eyes of the public. On the contrary, there are reasons

[12] The effect could be similar for people who are obese and receive conventional help from the healthcare system to lose weight. In cases where they fail (which is the most common outcome), they feel as if they are less in control than they were before they tried to lose weight, because they previously thought that they would succeed if they only applied themselves (see, e.g., Persson 2014).

to believe that transparency can actually harm public confidence or acceptance of decisions, not least because of the potential problems with fixing low quality algorithms. Of course, this is not to say that we should not give access to auditors, researchers, or any other actors who can decipher algorithms. Giving these groups access may yield higher-quality decisions and a higher degree of perceived legitimacy. Furthermore, there may also be good reason in favor of also demanding transparency in decision-making processes in the workings of AI when it comes to other agents, such as the decision-makers themselves. This is because they might use this information to make more well-informed decisions, thereby leading to higher-quality decisions that result in higher perceived legitimacy among the public. However, this does not imply that there are good reason in favor of sharing this information with the general public if perceived legitimacy is the goal.

4 The benefits of a justifications approach to AI transparency

In light of the previous discussion on the potential disadvantages of "transparency in process", we argue that a strategy which focuses on providing justifications for decisions (i.e., transparency in rationale) has the potential to generate perceived legitimacy among the public, both for decisions regarding the design of the AI in Phases 1 and 2, and for decisions made by the AI in Phase 3. In addition, we argue that a justifications approach can avoid many potential problems that arise from more demanding forms of transparency that bear the risk of backfiring on perceived legitimacy.

Generally speaking, a policy of justifications of decisions will inform the public of what the *decision* is, on which *grounds* it has been made, and in doing so, identify who the *responsible* actor is. In most cases, this information is sufficient for members of the public to form an opinion about the desirability of the decision and, if they so like, demand that the responsible actor be accountable (given, of course, that sufficient mechanisms for accountability are in place).[13] Thus, what we may want the decision-makers to do in Phases 1 and 2 is provide favorable reasons for the goals and priorities they have established for the AI and ensure that these reasons are made available to the public in a way that they can understand.

When it comes to Phase 3, the AI assistants should be able to provide explanations for their decisions or recommendations in an accessible language. It should also be clear who is accountable for the decision or where to turn if the individual wants to appeal the decision. Consequently, when AI assistants are helping a bank clerk to judge whether to grant a loan, the AI assistant could explain a negative decision by reporting that to get a loan of magnitude X, the customer needs to have collateral Y, because Z percent of people in the group below Y have defaulted in the past. Since the customer in question only has Y-100, she cannot receive the loan (cf. Wachter et al. 2017), and if she does not think that the decision is fair, she can argue her case by presenting it to the clerk, who then sets her appeal in motion.

There are reasons to believe that a requirement that provides justifications for decisions and policies in public can successfully ensure the legitimacy-enhancing mechanisms previously discussed. Roughly put, in Phases 1 and 2, the public eye monitors the reasons on which the decisions are based. Hence, the decision-makers will have incentives to behave better when it comes to producing such reasons, thus producing higher-quality decisions. Furthermore, the public will receive enough information to evaluate the desirability of the decision and, if they so like, demand accountability. Similarly, they are likely to get sufficient information to understand the different considerations and perspectives underlying the decision, thus making them feel included.

Delving deeper into the discussion, the first argument in favor of transparency in rationale is that according to principal–agent terminology, the agent (the people) will know who to hold accountable and for what, and this appears true for all phases. A policy of justification will tell the public what the decision is, on which grounds it has been made, and who the responsible actor is. In most cases, this information is enough for members of the public to form an opinion about the desirability of the decision and, if they so like, demand accountability of the responsible actor (given of course that sufficient mechanisms for accountability are in place). In many cases, a policy of justifications may even fulfill this role better than more demanding forms of transparency, e.g., when the whole process is made visible. Compared to more extensive forms of information provision, public justifications have the potential of being relatively short and condensed. This means that the public might have a better chance of actually finding and contemplating the relevant pieces of information than if they are provided with huge amounts of detailed information. As a policy, justifications might enable the public to adopt a strategy that, in the words of McCubbins and Schwartz (1984), is more like fire alarm oversight rather than police patrol oversight: they can act when something is obviously wrong but do not need to put a lot of effort on continuous monitoring. For a general public that is not particularly interested in political matters, this means that justifications may actually be a more attractive policy than full transparency.

Second, the decisions that the decision-makers make will probably be of higher quality because of what the decision-makers *need to do* and what they *can do* before facing the

[13] For a similar point, see Binns (2018,pp. 548–552).

public. When the process is complete and the final decision made, the decision-makers will need to explain themselves to the public. When decision-makers are required to provide reasons for their decisions, they are, in the words of Mill (1962, p. 214), forced "to determine, before he acts, what he shall say if called to account for his actions." Thus, decision-makers are likely to properly weigh the pros and cons when making decisions (Shapiro 1992), thus leading to higher-quality decisions. At the same time, when the actual decision-making process remains comparatively secret, decision-makers can ask important but seemingly "stupid" questions, change their minds in light of new arguments instead of negotiating on fixed preferences, and search for allies. These qualities can increase both quality and efficiency in decision-making (Mansbridge 2009). That decision-makers are given some shelter for discussion is especially important in Phases 1 and 2, since these phases are likely to contain both technically complex and morally controversial trade-offs. A cooperative working climate that can spark efficiency and problem-solving when it comes to AI in decision-making is crucial, as algorithms can potentially cause considerable damage if they are not properly adjusted when necessary.

An obvious objection is that focusing on the provision of justifications might increase the likelihood of decision-makers engaging in a post-decision construction of arguments designed to look better than they actually are—a maneuver known as window-dressing (Prat 2006). In other words, the justifications that are provided may not be genuine, and the only way to determine their sincerity is to access the process (Warren and Mansbridge 2013). This implies that transparency in rationale could yield worse decisions in term of quality, thus potentially decreasing perceived legitimacy. This reasoning is based on the concept that decision-makers act differently behind the scenes in comparison to how they act in public. Transparency is generally associated with the "myth of hidden politics" (Fenster 2006, p. 931; see also West and Sanders 2003). This myth is a public perception that the actual decision-making is something that takes place in smoke-filled rooms or private spaces that are hidden from the public arena, which is viewed as a scene of drama and spectacle. The widespread myth of secret power is sparked by the fact that modern popular culture generally tells us that we should be suspicious of anyone in a powerful position (Brin 1998).

Even though this is persuasive, it may be just a story. For example, Naurin (2007) has shown that the popular notion that powerful actors behave diametrically differently under conditions of transparency and conditions of non-transparency is not evident. He considers the example of European Union lobbyists, who can be expected to have incentives to behave in a civilized manner when it comes to political matters. Even if they act behind the curtains, they cannot say or suggest whatever comes to their minds. They must "get dressed for politics" (i.e., present serious and well-reasoned arguments for their positions). Taken together, this implies that the requirement to present justifications for decisions has a good chance of keeping decision-makers aware of their position as servants of the public. This will then lead to fairer decisions of higher quality, eventually leading to a higher degree of perceived legitimacy.

Third, there are reasons to believe that transparency in rationale could yield as much or even higher degrees of understanding. In Phases 1 and 2, when decision-makers are transparent about the decision and the reasons for taking it, the public should have greater understanding of why these decisions were made, since the information about the decision will be more condensed, thus making it more accessible. Opening the "black box" in Phase 3 in the sense that the AI gives explanations for its decisions rather than describing the whole process will also lead to greater understanding for the public. This is again because the public will get the information they need in a manageable way, meaning that they will not suffer from the information overload discussed in Sect. 3.

Finally, the public will likely view transparency in rationale as a fair way of dealing with these issues. To provide reasons for decisions signals that the decision-makers respect and care about the affected (e.g., Tyler and Lind 1992) and may lead to more favorable interpretations regarding the motives and intentions of the decision-makers (e.g., Bies and Shapiro 1988; Shapiro et al. 1994; Colquitt 2001; Shaw et al. 2003). Thereby, actors receiving justifications might be motivated to act as good losers, to accept that they have lost, and move on (e.g., Pitkin 1967; Gutmann and Thompson 1996). At the same time, when keeping the process more secret, we can avoid many of the potentially negative effects of full transparency previously discussed such as disappointment regarding the decision-makers' behavior in decision-making situations. Of course, exactly what is required for a justification to be perceived as sincere and adequate is a moot question that needs further examination. However, justifications need to be understandable and relevant to the decision at hand. These appear to be necessary conditions that make the receiver feel included and respected.

5 Conclusions

In this paper, we have produced a framework for analyzing transparency in AI decision-making in the socio-technological system and have argued that a limited type of transparency in the form of justifications for decisions—both regarding the design of AI assistants and the decisions taken by them—has the potential to ensure more legitimacy in the eyes of the public than transparency in process. When realizing perceived legitimacy, we should, as a default, opt

for having our AI assistants explain themselves rather than open up their code, etc. for public scrutiny. The same is true for the decisions of decisions-makers in the process when determining the goals and relevant considerations for the assistants.

It is evident that our analysis builds on a considerable number of assumptions and guesses. Thus, there is a need for both theoretical and empirical work that explores the role of justifications in decision-making in general, but specifically in relation to AI assistants. First, there is a need for a thorough analysis of what reasons should be normatively acceptable to use when publicly justifying decisions in a democratic setting. Second, there is a need for more empirical research regarding how justifications should be designed and presented to gain public acceptance. As argued by McGraw et al. (1995), several conditions must be met for an explanation/justification to have the intended effect: It must be exposed to the intended audience, the audience must pay attention to it, the audience must comprehend it, and the audience must accept the explanation/justification as legitimate. Third, there is a need to develop an empirically grounded theory for how a policy of justifications should be designed in practice to satisfy the demands of public insight and efficiency with regard to decision-making. Fourth, we need to evaluate how decisions and decision justifications are perceived by the public, depending on whether they are being made by human beings or AIs.

Our attempt with this paper has been to contribute to the discussion on transparency in AI decision-making using public perceptions that arose from making processes and justifications transparent. This discussion points to the importance of including a more thorough public perspective on AI design and decision-making that includes political decision-making and policy analysis as well as psychological insights into how individuals perceive authorities and authoritative decisions. However, more empirical and philosophical research must be done in this area before we have concrete knowledge on what to do and why.

Acknowledgements Open access funding provided by Chalmers University of Technology.

References

Binns R (2018) Algorithmic accountability and public reason. Philos Technol 31(4):543–556

Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) It's reducing a human being to a percentage: perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 Chi conference on human factors in computing systems, 1989 ACM. Bok S. Secrets: on the ethics of concealment and revelation. Vintage Books, New York, p. 377

Boscoe B (2019) Creating transparency in algorithmic processes. In: Blatt, et al. (eds) Jarbuch fur informationsfreiheit und Infomrationsrecht. Lexxicon Publisher, Berlin

Bostrom N (2017) Superintelligence: paths, dangers, strategies. Oxford University Press

Brin D (1998) The transparent society: will technology force us to choose between privacy and freedom?. Reading Addison-Wesley, MA

Carabantes M (2019) Black-box artificial intelligence: an epistemological and critical analysis. AI Soc. https://doi.org/10.1007/s00146-019-00888-w

Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (SP). IEEE, pp. 39 57

Chambers S (2004) Behind closed doors: publicity, secrecy, and the quality of deliberation. J Polit Philos 12(4):389–410

Chambers S (2005) Measuring publicity's effect: reconciling empirical research and normative theory. Acta Polit 40(2):255–266

Colquitt JA (2001) On the dimensionality of organizational justice: a construct validation of a measure. J Appl Psychol 86(3):386

de Fine Licht J (2014) Policy area as a potential moderator of transparency effects: an experiment. Public Adm Rev 74(3):361–371

de Fine Licht J, Naruin D, Esaiasson P, Gilljam M (2014) When does transparency generate legitimacy? Experimenting on a context-bound relationship. Governance 27:111–134

Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings. Proc Priv Enhancing Technol 2015(1):92–112

De Laat PB (2018) Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability? Philos Technol 31(4):525–541

Došilović, FK, Brčić M, Hlupić N (2018) Explainable Artificial Intelligence: A Survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 0210–0215

Elster J (1998) Deliberation and constitution making. In: Elster J (ed) Deliberative Democracy. Cambridge University Press, Cambridge

Eppler MJ, Mengis J (2004) The concept of overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. Inf Soc 20(5):325–344

Fearon J (1998) Deliberation and discussion. In: Elster J (ed) Deliberative democracy. Cambridge University Press, Cambridge

Fenster M (2006) The opacity of transparency. Iowa Law Rev 91(3):885–949

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Schafer B (2018) AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707

Grimmelikhuijsen S (2012) Linking transparency, knowledge and citizen trust in government: an experiment. Int Rev Adm Sci 78(1):50–73

Gunning D (2017) Explainable Artificial Intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web. Available here: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf. Accessed 13 Mar 2020

Holmström B (1979) Moral hazard and observability. Bell J Econ 10(1):74–91
Hood C (2006) Transparency in historical perspective. In: Hood C, Heald D (eds) Transparency: the key to better governance?. Oxford University Press, Oxford, pp 3–23
Hood C (2007) What happens when transparency meets blame-avoidance? Public Manag Rev 9:191–210
Hood C (2010) Accountability and transparency: siamese twins, matching parts, awkward couple? West Eur Polit 33(5):989–1009
Hosanagar K, Jair K (2018) We need transparency in algorithms, but too much can backfire. Harvard Business Review. Available here: https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire. Accessed 13 Mar 2020
Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2017) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 2017:1–17
Lindblom CE (1959) The science of 'muddling through'. Public Adm Rev 19(2):79–88
Manin B, Przeworski A, Stokes S (1999) Introduction. In: Przeworski A, Stokes S, Manin B (eds) Democracy, accountability, and representation. Cambridge University Press, Cambridge, pp 1–26
Mansbridge J (2009) A 'selection' model of political representation. J Polit Philos 17(4):369–398
McCubbins MD, Schwartz T (1984) Congressional oversight overlooked—police patrols versus fire alarms. Am J Polit Sci 28(1):165–179
McGraw KM, Best S, Timpone R (1995) What they say or what they do? The impact of elite explanation and policy outcomes on public opinion. Am J Polit Sci 39(1):53–74
Meade M, Stasavage D (2006) Two effects of transparency in the quality of deliberation. Swiss Polit Sci Rev 12(3):123–133
Mill JS (1962) Considerations on representative government (1861). Gateway, South Bend, Indiana.
Naurin D (2007) Deliberation behind closed doors: transparency and lobbying in the European Union. ECPR Press, Colchester
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Broadway Books, New York
Pasquale F (2015) The Black Box Society: the secret algorithms that control money and information. Harvard University Press, Cambridge
Pitkin H (1967) The concept of representation. University of California Press, Berkeley
Prat A (2006) The more closely we are watched, the better we behave? In: Hood C, Heald D (eds) Transparency: the key to better governance?. Oxford University Press, Oxford, pp 91–103
Richards NM, King JH (2013) Three paradoxes of big data. Stan L Rev Online 66:41
Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach. Pearson Education Limited, Malaysia
Scharpf F (1999) Governing in Europe: effective and democratic?. Oxford University Press, Oxford
Shapiro M (1992) The giving reasons requirement. U Chi Legal F 1992:179
Shaw JC, Wild E, Colquitt JA (2003) To justify or excuse?: a meta-analytic review of the effects of explanations. J Appl Psychol 88(3):444–458
Sweeney L (2013) Discrimination in online ad delivery. ACM Queue 11(3):10
Thelisson E, Padh K, Celis LE (2017) Regulatory mechanisms and algorithms towards trust in AI/ML. In: Proceedings of the IJCAI 2017 workshop on explainable artificial intelligence (XAI), Melbourne, Australia
Tyler TR (2010) Legitimacy and rule adherence: a psychological perspective on the antecedents and consequences of legitimacy. In: Bobocel DR, Kay AC, Zanna MP, Olson JM (eds) The psychology of justice and legitimacy. Taylor and Francis, New York
Tyler T, Lind EA (1992) A relational method of authority in groups. Adv Exp Soc Psychol 25:115–191
Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76–99
Warren ME, Mansbridge J (2013) Deliberative Negotiation. In: Mansbridge J, Martin CJ (eds) Negotiating agreement in politics: report of the task force on negotiating in politics. American Political Science Association, Washington
West HG, Sanders T (2003) Transparency and conspiracy: ethnographies of suspicion in the new world order. Duke University Press, Durham
Zednik C (2019) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol. https://doi.org/10.1007/s13347-019-00382-7
Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci Technol Human Values 41(1):118–132
Zerilli J, Knott A, Maclaurin J, Gavaghan C (2018) Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 32(4):661–683